DOCUMENT RESUME

ED 300 415 TM 012 411

AUTHOR Chipman, Susan F.
TITLE Word Problems: Where Test Bias Creeps In.
PUB DATE Apr 88
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988). Portions of appendices may be marginally legible.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Arithmetic; Elementary Secondary Education; *Item Analysis; *Mathematics Tests; *Problem Solving; Semantics; *Sex Bias; *Test Bias; Testing Problems; *Word Problems (Mathematics)

ABSTRACT

The problem of sex bias in mathematics word problems is discussed, with references to the appropriate literature. Word problems are assessed via cognitive science analysis of word problem solving. It has been suggested that five basic semantic relations are adequate to classify nearly all story problems, namely, change, combine, compare, vary, and transform relations. The idea that rapid penetration to formal mathematical structure of a problem is the essence of the word problem solution is contrasted to the often sexually-biased psychological reality faced by the examinee. Research on logical reasoning has largely neglected the influence of problem content, which is often sexually biased, on the solution of mathematics word problems. Nevertheless, there is a great deal of research that indicates that problem content has a great influence on performance. Experimental studies on the effects of problem content, incorporating possible sexual biases, are recommended. Four graphs and two tables showing masculine, feminine, and neutral test items conclude the document. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300415

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
SUSAN F. CHIPMAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Word Problems: Where Test Bias Creeps In

Susan F. Chipman

April, 1988 Paper presented at AERA, New Orleans

Popular belief greatly exaggerates the extent of sex differences in performance on tests of mathematical ability or skill. Sex differences in mathematics achievement are negligible prior to high school and parallel the relatively small sex differences in mathematics course enrollments in high school (Chipman & Thomas, 1985). For example, the 1982 National Assessment of Educational Progress mathematics assessment showed no sex difference in overall mathematics scores at ages 9 and 13 and a difference of less than 2 points (on a scale with a mean near 50) at age 17 (NAEP, 1983). Additional confirmation of such results comes from a recent analysis of the California Assessment Program data for sixth grade; even an analysis which broke the results down by item type found only a small significant difference in favor of girls in computation items (Marshall & Smith, in press). When sex differences that favor males are found, however, they tend to involve tests that are composed of word problems. For example, in the Second NAEP Mathematics Assessment (Armstrong 1979, 1985), 13 year old females performed better on computation items (53.4% vs. 49.9%) and showed a statistically insignificant tendency to perform better on algebra items

TM 012 411

BEST COPY AVAILABLE

(53% vs. 51%) whereas males performed slightly better on the problem solving items (44.1% vs. 42.5%). At age 17, the only statistically significant difference was on the problem solving items, where males performed better (46.1% vs. 41.0%). Unfortunately such differences tend to be relabeled as differences in "mathematical reasoning ability" or in the "higher cognitive levels" of mathematics.

Therefore, it seems to be worthwhile to look below the surface of these tests, to look at the individual word problem items and the thought processes involved in solving them, to consider where sex differences might arise in those thought processes, and to consider how such differences really ought to be interpreted when they are observed.

Cognitive Science Analyses of Word Problem Solving

By now, both cognitive scientists and more traditional mathematics educators have devoted considerable research attention to the processes of solving word problems. Early on, Hinsley, Hayes & Simon (1977) and Paige & Simon (1966) sketched the processes of solution. Within the reading of the first phrase of a problem, subjects may recognize it as an instance of a typical category of problems -- such as "river current problems", "mixture problems" or "work problems". Recognition of the problem category is

associated with other information such as expectations about the question that will be posed and about the equations and types of drawings that may be useful in solving the problem. Strategies for isolating the relevant information from the problem statement and for disregarding irrelevant information are also invoked. There may also be affectively toned expectations that this class of problems is difficult and that failure is likely.

Mayer (1981) further developed and differentiated the description of word problem categories with a thorough analysis of the problems appearing in a set of textbooks. He noted also that texts often explicitly name these categories and provide methods of solution for them (see especially Rich, 1973). Empirical studies of problem categorization and perceived similarity were conducted by Silver (1979) and Chartoff (1977). These will be discussed further below, but they do seem to provide some confirmation for the salience of categories of the type described by Hinsley, Hayes & Simon (1977) and Mayer (1981): although it was not anticipated in the design of his study, a dimension which he labeled "pseudostructure" -- because it was not purely mathematical structure -- emerged in Silver's data. One of the "dimensions" of similarity that Chartoff (1977) reported was "comparison with a generic problem of the same type".

Hinsley, Hayes & Simon (1977) demonstrated the

importance of problem categories in several ways. Problems with the same mathematical structure may be solved in quite different ways if the cover stories cause them to be categorized differently. Subjects will even approach the solution of problems that are posed in nonsense words by attempting to assimilate them to one of these categories. Sometimes however, when subjects are unable to categorize a problem, they will resort to a fallback strategy which appears to be a relatively direct sentence to sentence translation from the words of the problem to mathematical expressions. At first consideration, it may seem that the problem solver's task begins with the translation of words into mathematical terms. An early computer simulation,STUDENT, (Bobrow, 1968) emphasized this sort of translation process, but Paige & Simon (1966) observed that human problem solvers show more use of semantic information about the content of the problem even when they are using a translation procedure. Individuals also vary in the degree to which they tend to use translation strategies.

Mathematics educators often complain that students use strategies dependent upon key words to select operations and employ heuristics like being certain to use every number that appears in the problem, without really understanding the problem. Often -- given the way textbook problems are written -- the strategies work, and some teachers even teach them. Unfortunately, the information

that the real world provides about situations in which it might be appropriate to use mathematics is not so nicely tailored to fit these strategies. Deeper understanding is necessary, and cognitive science research is beginning to explicate what it means to genuinely understand problems.

Arithmetic Word Problems: Models for the Simpler Case

Drawing upon a general theory of text comprehension (van Dijk & Kintsch, 1983) and hypotheses about the semantic knowledge needed for understanding word problems (Riley, Greeno & Heller, 1983), Kintsch & Greeno (1985) have made a somewhat more detailed proposal about the processes involved in comprehending and solving arithmetic word problems. The memory representation produced as the result of reading a text is thought to consist of two parts -- a propositional representation of the information that is in the text and a model of the situation described by the text. The situation model is derived both from information in the text and from inferences that can be made from that information using one's general knowledge of the topic area of the text. Marshall, Pribe & Smith (1987) point out that even simple arithmetic word problems typically require inferences based on superordinate-subordinate class membership knowledge that the solver provides from memory. Paige and Simon (1966) provided an analysis of the physical conservation assumptions and inferences based upon them that were

necessary to make sense of a liquid mixture problem they studied. Thus, the task of understanding a word problem is to construct a conceptual model of the problem situation on which problem solving operations can operate. Kintsch & Greeno (1985) point out that there are special strategies for comprehending word problem texts and constructing models of the situations they describe. There are special presuppositions in word problem comprehension that do not apply to ordinary texts; one ignores many of the features of objects and people that would ordinarily be important in story comprehension. As Marshall, Pribe and Smith (1987) put it, when one reads a problem like this, "Mary had 5 oranges and 3 apples. How many pieces of fruit did she have?" one does not speculate about how many bananas and apricots she might also have. Certainly the motives that characters might have for "giving" and "taking" are not a matter of concern. The goal is simply to identify the sets of objects and the relations among them that are the basis for selecting appropriate arithmetic operations.

It appears that Marshall (Marshall, 1985, 1987; Marshall, Pribe & Smith, 1987) has made the most comprehensive proposal about the class of prototypic underlying situations (or schemata) that are described in arithmetic word problems. She has found that 5 semantic relations are adequate to classify nearly all of the story problems appearing in sixth grade texts used in California,

in community college remedial math materials, and in Navy remedial math materials. (The remaining problems typically required the application of geometry or probability formulas with little semantic interpretation.) The relations she proposes are the following:

1) The Change relation. The relation is an increment or decrement of a measurable resource. There is an initial state and a final state, and these cannot exist at the same time. Most involve additive change (thus require the use of addition or subtraction), but the situation could be one of multiplicative or exponential growth.

2) The Combine relation. It involves the renaming of elements with respect to a superordinate category. The subordinate categories must have semantic ties to the superordinate one (apples, orange, fruit), and usually the problem solver must supply that semantic information from long term memory in addition to having an understanding of subordination relationships.

3) The Compare relation. Two elements of the problem are evaluated in order to determine their relative size. They must have the same semantic features, such as the unit of measurement.

4) The Vary relation. The problem structure involves 4 quantities -- such as a single object with an associated unit price and a group of those objects with an associated total price.

5) The Transform relation. Its essential feature is the understanding that one object can be described in several ways, as in this problem: "Sue is 1/3 as old as her mother. If her mother is 30 years old, how old is Sue?" These problems involve two distinct entities described in the same object-units (eg. years) with a mathematical relation between those descriptions.

These semantic relations are very generic situation descriptions or models. Several different questions can be asked to pose problems about them, depending upon which information is given about the situation and which is to be found. For this reason, and because a variety of different mathematical relations are possible within these situations (cf. the description of the Change relation), identifying the underlying semantic relation is not the same as identifying the mathematical operation to be used in solving the problem. More complex problems requiring multiple steps for their solution can be built by combining these relations, but they rarely occur in textbooks and training materials.

The computer simulation models of problem solving that Marshall is developing proceed by attempting to fit these semantic relations to the problems, testing the necessary conditions or constraints that must be met. This represents the type of strategies for identifying relevant and irrelevant information that Simon and his associates

described. Once a semantic relation is identified, this together with the known and unknown information, guide the selection of appropriate arithmetic operations. Each aspect of the simulation, of course, involves a great deal of detailed knowledge that it is impossible to discuss here. For example, if some of the constraints required by a semantic relation are disregarded, erroneous problem solving behavior can result. Although we are largely unaware of this level of detail when we solve problems, presumably all of this knowledge must somehow be acquired by successful students.

It is probably appropriate to regard the "semantic relations" proposed for arithmetic word problems as a small subset of the problem categories that exist for algebra word problems. Mayer (1981) identified numerous problem categories and subtypes or "templates" within those associated with the different questions that might be asked. It is possible that certain highly practiced solvers of word problems may know all or most of these types, enabling them to rely upon the top-down schema-fitting strategy that Marshall suggests for arithmetic problems. Most solvers, however, must rely upon some more general strategies for identifying underlying quantities and the constraints that hold among them. For example, the strategy of drawing diagrams to represent the physical situation described in a problem is a strategy that one

often sees early in the solution attempt (Webb, 1979).

Mathematical Idealism and Psychological Reality

Mathematics educators consider that rapid penetration to the formal mathematical structure of a problem is a key manisfestation of matnematical ability. The Russian psychologist Krutetskii (1976) has said that this is one of the characteristics of mathematically gifted students, and this observation inspired Silver's (1979) study of problem categorization that was mentioned above. In the set of problems that he presented to an above average group of eighth graders, Silver contrasted "mathematical structure" to "contextual details" as possible bases of categorization. It turned out that students also classified problems on the basis of the type of question asked and on the basis of "pseudostructure" -- age problems, time problems, weight problems, etc. Therefore, he did a second study in which the 4 variables -- mathematical structure, pseudostructure, question type, and contextual details -- were more cleanly separated from each other. He did find that a tendency to categorize by mathematical structure was associated with good problem solving performance, and, not surprisingly -- especially since "correct" solutions were presented in class in the interim -- that classification by mathematical structure increased after the experience of solving the problems. But "pseudostructural" classification also came out looking

reasonably good, whereas classification according to contextual details characterized poorly performing students. It is not entirely clear what is meant by "mathematical structure" in this study -- in the first experiment, the classification of problems was determined by the investigator; in the second, it was determined by the investigator in consensus with 5 other mathematics educators. Probably it is something like, "how the problems are solved", which was one of the "dimensions" Chartoff (1977) found in similarity judgements of problems. Or perhaps it is more accurate to say that it has to do with the mathematical equations that the investigator would choose to set up in solving the problems because the discussion comments indicate that many of the students, who had not yet been taught algebra, solved the problems successfully by successive approximation. It seems likely, therefore, that Silver's "mathematical structure" is much like the semantic relations or situation models that Marshall, Kintsch, and Greeno have discussed. For the more complex case of algebra problems, the abstract equivalent seems to be the quantities and constraints among them that exist in the problem. Various different solution approaches might be consistent with those relations. Because "mathematical structure" has not been formally and precisely defined in these studies, it is difficult to know exactly what it was, and the de facto definition may not

have been entirely consistent. Clearly, it is at least the problem categories that are perceived by relatively more expert solvers: mathematics educators rather than students.

Close examination of what people are doing when they solve word problems can be a surprising experience; "the way that the problem is solved" can dissolve into great diversity. Instant penetration to "mathematical structure" certainly appears rare. Hall, Kibler, Wenger, and Truxaw (1986) recently collected such data in preparation for a computer simulation of algebra word problem solving. They collected written solution protocols on four problems from 85 students who could be expected to be relatively competent problem solvers: 85 juniors and seniors in computer science at the University of California, Irvine, where admission is somewhat selective and where these students had to have completed at least three university level courses in calculus. They found a great deal of what they called "model-based reasoning" with the situational context of the problem (22 to 47% of the subjects, depending upon the problem) and relatively little formal algebraic reasoning although most used some algebra (64 to 86%). Although the problems consisted of two mathematically isomorphic pairs of motion and work problems, the motion cover stories inspired the use of diagrams (69 and 37%) whereas the work problems generally did not (8 and 9%). Given this difference in solution

processes, it is not surprising that success rates were very different for the two members of the mathematically isomorphic pairs: (91% vs. 61%) and (52% vs. 92%). In one pair, the motion problem was easier; in the other, the work problem. Thus, content can make a very big difference in problems with the same mathematical structure, and people often do considerable work rather directly with the problem content in solving algebra word problems. The picture of problem solving performance that emerged from these data was considerably more complex than these researchers anticipated when they set out to simulate it, just as it seems to be much more complex than most mathematics educators and test-builders imagine. The early paper by Paige & Simon (1966) also provided striking demonstrations of model-based reasoning in doing word problems. One of their word problems described a physically impossible situation in the instructions for cutting board lengths. Although some subjects did simply write down the mathematical relationships, others immediately noted the impossibility, and still others transformed the problem as given into one that was physically possible. Similarly, the drawing of integrated diagrams representing the liquid mixture problem mentioned above seemed important in achieving a successful solution.

The cognitive science analyses tend to demystify such things as "seeing mathematical structure", to show what

that means and how it is done, when and if it is done. They imply a very complex structure of acquired knowledge -- categories and subcategories of problems, knowledge of the conditions that need to be met for particular problem types, knowledge of strategies for representing problem information effectively in diagrams, knowledge of good strategies for proceeding with a given problem type, etc. Of course, in combination with cognitive theories of complex learning (e.g. Anderson, 1983), they imply that these categories and skills may become much more consolidated, efficient, and reliable with extensive practice. This view of what is going on in problem solving performance is very different from the simple psychometric model of a single dimension of "mathematical reasoning ability".

Related Research in Logical Reasoning

There is an obvious, striking resemblance between this research on the solving of word problems and research on logical reasoning -- but there is much more of the latter in psychology. Here, as in mathematics, there has been a feeling that people ought to think like formal logic, but much evidence that what they are doing even when they successfully perform on problems in logic has little resemblance to formal logic. Johnson-Laird (1983, 1985) has proposed that people do logic problems and inference by building "mental models" that represent the situation

described in the problem -- for example, imagining sets with the specified characteristics. So, if we are told that "some men are mortal", we might imagine a group of men wearing tee-shirts, some of the tee-shirts having the word "mortal" on them. (Or, the model could be more abstract and reduced than this suggests.) When we are told Socrates is a man, then we check our image to see whether it tells us anything about whether or not Socrates is mortal. Problems become difficult when there are substantially different situations that could meet the constraints mentioned in the problem and one needs to imagine and consider all of them in order to ensure a correct response. Of course, people can be more or less skilled in thinking of and building all of the necessary model cases.

Whereas there seems to be little or no research on the influence of problem content on the solving of mathematics word problems, there has been considerable research on the influence of problem content on performance in logical reasoning tasks. Content makes a big difference. In general, people have difficulty in letting the logic game override what they know about real life. Since you know that all men are mortal, you are likely to conclude that "Socrates is mortal" even though I gave you the premise that "Some men are mortal". Embodying an inference problem in a real-life situation that represents it correctly makes it much easier. Wason and Johnson-Laird (Johnson-Laird,

1983) have given simple and dramatic demonstrations of this. Given a set of cards with a letter on one side and a number on the other and the instruction to verify this generalization, "If a card has a vowel on one side, then it has an even number on the other side," subjects have great difficulty correctly deciding which of these cards needs to be turned over for examination: E,K,4,7. But if the set is (Manchester, Sheffield, Train, Car) and the rule is, "Every time I go to Manchester I travel by train," performance is much better. Similar good results are obtained with a rule like, "If a letter is sealed, then it has a 50 lire stamp on it."

The message of this for the solving of mathematics problems is certainly that the familiarity of problem content could make a big difference. In the small set of problems investigated by Hall et. al. (1986), content certainly did make a big difference -- although not necessarily for reasons of familiarity. Because the situations that may appear in word problems are life situations which may differ in familiarity for males and females, this is one possible source of sex differences in performance. In order to understand a problem, inferences may need to be made that depend upon one's knowledge of the content described in the problem. Building up and holding in mind a model of the situation described in the problem is likely to be easier with familiar content. It may even

be true that the experiences of males and females provide differential practice with some types of problems -- such as the computation of sports averages.

What if anything do we know about the relation of problem content to sex differences in performance?

Problem Content and Item Bias

There have been a number of studies which have attempted to investigate, after the fact of test construction and administration, whether some items seem to be biased for or against females or males and whether something systematic can be said about the character of those items. Donlon (1973) studied the performance of the entire population taking the May, 1964 SAT. On the average, the 60 math items slightly favored males. There were 17 items with real-world referents (e.g. pulleys, cars, wheels, etc.) and 19 items that could be called algebra items. The content of the real-world problems was masculine in character, as the examples indicate, and there were no female actors in the problems. The only item of this type that favored females involved a household setting. On the average, females performed relatively better on the algebra items. Also, females were relatively less disadvantaged on a type of item which called for judging whether sufficient information was available to solve the problem, without actual solution. A later study

of the April, 1974 SAT (Strassberg-Rosenberg & Donlon, 1975) confirmed that data sufficiency items were somewhat more favorable to females. Geometry regular math items and regular math subject matter items were relatively most favorable to males. The content of the subject matter items appeared to be less sex-stereotyped than it had been in 1964. There were 8 female references as compared to 18 male references, and six items were judged to involve male interests or skills, 4 female interests and 2 neutral. No particular relationship could be seen between individual item bias and stereotypic content; however, the number of problems is very small and the number of variables that might be relevant is large. As the figure (Figure 1) taken from this paper shows, the effects under discussion are rather small and unlikely to be sufficiently reliable to support conclusions about item types under these circumstances.

In another study of the December, 1974, Graduate Record Examination, Donlon, Hicks, & Wallmark (1980) found 9 items which seemed to show some evidence of bias in favor of either males or females. Three of the items which favored males all involved the interpretation of a single economics graph. There were two geometry items which appeared to be biased in favor of females and one that appeared to be biased in favor of males. One algebra item and one miscellaneous item appeared to favor females. Four

of the 5 items favoring females called for finding algebraic unknowns. Again, it is difficult to make much of such results, and the figure (Figure 2) from this paper shows that the "outliers" which are being discussed do not lie extremely far out.

In an analysis of the results of the second NAEP mathematics assessment, Fennema & Carpenter (1981) reported that geometry and measurement items tended to favor males, but they were unable to find any consistent relationship between the apparent demands on spatial skills and the magnitude of the sex difference. Perhaps other content variables that they did not examine were more important. In any case, although the first hint of a sex difference on geometry items tends to evoke explanations in terms of the allegedly well-known sex difference in spatial skills, tests of most kinds of spatial ability do not show sex differences (Chipman & Wilson, 1985; Linn & Petersen, 1985).

It is not surprising that yet another study of this kind, Becker (1983), found a different pattern of performance among the males and females taking the SAT as part of the Johns Hopkins Study of Mathematically Precocious Youth. Becker also found the data sufficiency items to be easier for girls. Unlike Donlon and Fennema & Carpenter, however, she found algebra items to be more difficult for girls (perhaps because they were

predominantly word problem -- that is suggested by the discussion). There was no evident sex difference on geometry or arithmetic. Becker also found a serial position effect on difficulty that was greater for girls, suggesting that they might be working more slowly and cautiously.

The Need for Experimental Evidence

Post-hoc analyses of the items that happen to have been included in existing tests suffer from many limitations. The number of relevant problems is small and the number of possible variables affecting their difficulty is large. It is not surprising, therefore, that the results of such analyses have shown the inconsistency described above.

The discussion of possible sex bias in test items suffers from an extreme lack of direct experimental evidence concerning the effects of the variables that are thought to be important. One much-cited paper (Graf & Riddell, 1972) studied only 2 problems, one considered to have masculine content and the other feminine, and only one of these problems was administered to each subject in a small sample of college students. Reputedly, Milton (1958), as cited in Graf & Riddell (1972) and elsewhere, found that when the content of problems was made relevant to the female role -- cooking recipes, shopping, etc. -- males no longer showed superior performance, but as yet I have been unable to obtain a detailed report of this study.

For this reason, as part of joint project with David Krantz of Columbia and Rae Silver of Barnard, I took the opportunity to insert matched sets of problems in a mathematics placement test being administered to entering Barnard students. The idea was to create mathematically identical problems on alternate test forms that differed only by having a stereotypically masculine or feminine cover story (as judged subjectively by us). Thus, a mixture problem appeared in the guise of either making lemonade or mixing oil and gasoline to fuel an outboard motor. Changing the size of a beanbag chair to alter the amount of stuffing needed was paired with changing the size of an oxygen tank on a NASA space mission; a work problem appeared as shared effort hemming drapes or doing crop dusting; the trade-off in cost of two dairy products was paired with trade-off in gasoline consumed by low and high altitude flying. Unfortunately, it was not possible to insert very many problems of this type into the test.

The results (Table 1) did not show a consistent effect of the sex-typing of the cover story. The lemonade problem and the dairy product problem were both easier than their counterparts, but the hemming and crop dusting problems did not differ. Nor did the beanbag chair and NASA tank problems differ, perhaps because the problems were too difficult, really at the chance level of performance. It is difficult to know what to make of this pattern of

results, especially since no comparable data are available from a male population. For one thing, the general familiarity or typicality of the problems differs a great deal also. (Algebraic and numeric versions of the problems were compared as well. On every one of these problemsexcept the difficult beanbag/NASA problem, and also on an additional problem which provided only the algebraic vs. numeric comparison, the algebraic version proved more difficult.)

Sex-typed Problem Content Does Make a Difference

There seemed to be a need for a larger sample of problems and more careful consideration of possible relevant variables. Carol Tittle drew my attention to a dissertation (McCarthy, 1976) in which a large number of word problems (120) had been rated for relative familiarity of content to males and females by a group of 50 high school students. McCarthy's summary of the factors determining the ratings she got in a pilot study is of some interest. She found good agreement between males and females. She found that if only the individuals named or pronouns were sex-linked, the item would be considered neutral. Female items dealt with traditionally female skills such as typing, organizing picnics, cooking and sewing. Masculine items included sports and building activities. Given this insight into the determinants of perceived sex-typing, the 120 items that were rated were

constructed to consist of sets of 40 masculine, 40 feminine and 40 neutral items with comparable difficulties and skill requirements, as judged by teachers. Once rated, the items were used in a study of the impact of standard test construction procedures in producing possibly sex-biased tests. As part of that, a 78 item test consisting of 26 strongly masculine items, 26 strongly feminine items, and 26 neutral items was administered to 185 male and 148 female high school students. Together, the students doing the ratings and the students tested comprised the entire 10-12th grade enrollment in academic (i.e. college-bound) mathematics in a large suburban high school in the New York area.

A difficulty parameter, p, the percentage of students getting the item correct, was reported separately for males and females for each of the items in the 78 item test, but these values were not presented together in a single table nor were they directly analyzed in relation to the item characteristics. Therefore, I decided to undertake that analysis, and the results amazed me. The figures show the female p value on the ordinate plotted against the male p value. The first (Figure 3) is for the neutral items only. It resembles the figures from the ETS studies and shows slightly better male performance. The second (Figure 4) is for the items that were rated as masculine. Every single one of them falls below the 45-degree line of

equality -- many far below. The third (Figure 5) shows the items rated as feminine -- virtually all above the line. If ever a data set passed the interocular trauma test of significance this is it. The sex-typing of item content made a whopping difference to student performance, for both females and males. The items that are the extreme outliers -- most biased in favor of females or males -- are shown in Table 2. They are certainly sex-typed, but they do not require arcane knowledge unlikely to be available to students of the opposite sex. In this, they resemble the sports average item that showed a large sex difference in the Princeton Review data (cf. Loewen, Rosser, & Katzman, 1988). It is evident that the selection of item content could make quite a difference to the average test scores obtained by males and females.

Explaining Content Effects

How can these striking effects of sex-typed problem content be understood?

Somewhat outside of the cognitive framework discussed above, the explanation could be a purely affective avoidance of subject-matter that is considered to be owned by the opposite sex. Probably this hypothesis would be attractive to many, and if it is correct, it signals a serious psychological problem needing to be addressed. If this is so, it would suggest that (some) students simply omit and pass by the items that they consider to have

content appropriate for students of the opposite sex. The p values that index item difficulty are sometimes computed using only the data for items attempted or reached, but there is no indication that McCarthy (1976) did that. Thus, "not correct" could mean that the item was omitted.

In collecting new data, it would, of course, be easy to determine whether students were omitting items. But there are other reasons why items might be omitted. Students might feel that they do not know enough about the situation described in a problem to be successful in solving it, especially under the time pressure of a test. Note that, in the cognitive science view, the early stages of problem solving involve building a "situation model", and that drawing inferences for that purpose may depend upon one's general knowledge of the problem content. A related possibility is that the specific content may lead to a categorization of the problem as being of a type for which the student does not have a readily available, highly practiced strategy. It is not obvious how one would empirically differentiate these possible reasons for item omission.

Suppose we were to find that the problems are attempted, then how might the differential familiarity of content make a difference in performance? In a few cases, students may have highly practiced routines for particular problem content, such as sports averages. Extreme interest

in, and associated practice in computing, sports averages is quite common among young boys, rare among young girls. Perhaps this explains why many of the items showing extreme bias in favor of males involve the computation of sports averages. In general, cognitive science research has shown that transfer of problem solving strategies is highly dependent upon surface features of the problem (Polson & Jeffries, 1985) and that the perception of analogies between situation depends heavily upon surface features even though the judged quality of analogies depends upon deeper structural features (Gentner,).

Problem familiarity may impact on the ease of building situation models or the ease of making inferences even when all the necessary information seems to be there and accessible to everyone. The work of Wason and Johnson-Laird seems to demonstrate that clearly. Familiarity of content is generally associated with building "chunked" knowledge structures that make it easier to organize and reliably remember problem information while one also works with it. Consequently, it may be worthwhile to investigate whether content familiarity alone can account for differential problem success, regardless of the relation to sex roles. Such investigations would be valuable in contributing to the understanding of other possible forms of cultural bias in tests of mathematical skill or ability. It is likely that the content of

different problems differs in familiarity for members of different ethnic groups and social classes.

Although mathematics educators may see their goal as developing insight into the pure structure of situations that can be represented by mathematics -- or may see the ability to penetrate to such structure as characteristic of mathematically gifted individuals -- there may be an element of myth in this goal. One may simply observe the result of extensive practice (and perhaps individual efficiency differences in the building of knowledge structures) so that deep and general categories have been formed and can be rapidly accessed via more surface properties. And it seems that few individuals ever attain that level of performance.

If that is the goal, it may pose interesting new challenges to mathematics educators and suggest new approaches. One might try to teach the recognition of underlying structures in relatively direct ways. Sandra Marshall is in fact trying to do that for arithmetic word problems. This may be possible. But it seems likely that the structures for algebra word problems may be so much more varied and flexible that a different conception of the goal of instruction is needed. Perhaps students must learn procedures for constructing situation models or mathematical structures, not merely criteria for identifying one of a small set of standard structures. It

may help -- if one can communicate the idea -- to make the search for quantities and the constraints that relate them an explicit goal. A recent study of transfer of procedures for doing series problems done by Bassok & Holyoke (1987) provides somewhat encouraging evidence that mathematics courses do succeed somewhat in conveying the message that problem content is not the important issue. Training in the context of mathematics provided better transfer than training in the context of physics and better transfer than has been observed in most cognitive science studies of transfer.

It was noted above that there is quite a contrast between the cognitive science view of what goes on behind expert performance in solving word problems and some simple notion of (often presumed to be innate) "mathematical reasoning ability". The latter view tends to disregard or deemphasize the effect of specific practice and experience. More and more, we are beginning to recognize that Americans have a tendency to emphasize "ability" as an explanation of mathematics performance in a way that is probably dysfunctional. These beliefs almost certainly contribute to generally low levels of mathematics performance as contrasted to what one observes in other countries. Recognizing that there is a great deal to be taught, and a great deal to be learned, would improve mathematics education for everyone in the U.S., male and female, majority and minority.

REFERENCES

Anderson, J.R. (1983), The Architecture of Cognition. Cambridge, MA: Harvard University Press.

Becker, B. J. (1983), Item characteristics and sex differences on the SAT-M for mathematically able youths. Paper presented at AERA, Montreal, April, 1983.

Bobrow, D.G. (1964), A question-answering system for high school algebra word problems, AFIPS Conference Proceedings, 26. Fall Joint Computer Conference. Baltimore: Spartan Books.

Bassok, M. and Holyoak, K.J. (1987), Interdomain transfer between isomorphic topics in algegra and physics. Unpublished manuscript, LRDC, University of Pittsburgh.

Chartoff, B.T. (1977), An exploratory investigation utilizing a multidimensional scaling procedure to discover classification criteria for algebra word problems used by students in grades 7-13. (Doctoral dissertation, Northwestern University, 1976). Dissertation Abstracts International, 37, 7006A.

Chipman, S.F. Brush, L., & Wilson, D. (Eds.). (1985). Women and mathematics: Balancing the equation. Hillsdale, NJ: Erlbaum.

Chipman, S.F., Segal, J.W. and Glaser, R. (1985), Thinking and Learning Skills. Volume 2: Research and Open Questions, Hillsdale, NJ: Erlbaum.

Chipman, S.F. & Thomas V.G. (1985). Outlining the problem. In S.F. Chipman, L. Brush, & D. Wilson (Eds.), Women and mathematics (pp. 1-24). Hillsdale, NJ: Erlbaum.

Chipman, S.F. & Wilson, D. (1985). Understanding mathematics course enrollment and mathematics achievement: A synthesis of the research. In S. F. Chipman, L. Brush, & D. Wilson (Eds.), Women and mathematics (pp. 275-328). Hillsdale, NJ: Erlbaum.

Donlon, T.F. (1973). Content factors in sex differences on test questions (ETS RM-73-28). Princeton, NJ: Educational Testing Service.

Donlon, T.F., Hicks, M.M., J. Walmark, M.M. (1980). Sex differences in item responses on the graduate record examination. Applied Psychological Measurement, 4, 9-20.

Fennema, E. & Carpenter, T.P. (1981). Sex-related differences in mathematics: Results from National Assessment. Mathematics Teacher, 74,554-559.

Gentner, D. ( ) The mechanisms of analogical reasoning. In S. Vosniadou & A. Ortony (Eds.). Similarity and Analogical Reasoning. New York: Cambridge University Press.

Graf, R.G., Riddel, J.C. (1972). Sex differences in problem-solving as a function of problem content. The Journal of Educational Research, 65, 451-452.

Hall, R., Kibler, D., Wenger, E. & Truxow, C. (1986). Exploring the episodic structure of algebra story problem-solving. Technical Report 86-24. Department of Information and Computer Science, University of California, Irvine.

Hinsley, D.A., Hayes, J.R. & Simon, H.A. (1988). From words to equations: Meaning and representation in algebra word problems. In M.A. Just & P.A. Carpenter (Eds.). Cognitive processes in comprehension. Hillsdale, NJ: Lawrence Erlbaum Associates.

P. N. Johnson-Laird (1983). Mental Models. Cambridge, MA: Harvard University Press.

P. N. Johnson-Laird, (1985). Logical Thinking: Does it occur in daily life? Can it be taught? In: Chipman, Segal, and Glaser (Eds.) (p. 293-318).

Kintsch, W. & Greeno, J.G. (1985). Understanding and solving word arithmetic problems. Psychological Review, 92, 109-129.

Krutetskii, V.A. (1976), The Psychology of Mathematical Abilities in Schoolchildren. Chicago: The University of Chicago, Press. (original Russian edition. 1968).

Linn, M.C. & Petersen, A.C. (1985). Emergence and characterization of sex differences in spatical ability: A meta-analysis. Child Development, 56, 1479-1498.

Marshall, S.P. (1985). An analysis of problem-solving instruction paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Marshall, S.P. (1987). Schema knowledge for arithmetic story problems: Some issues of instruction, cognition and application. Unpublished manuscript, Department of Psychology, San Diego State University.

Marshall, S.P., Pribe, C.A., & Smith, J.D. (1987). Schema knowledge structures for representing and understanding arithmetic story problems. Technical Report, Department of Psychology, San Diego State University, March, 1987.

Marshall, S.P., & Smith, J.D. (1987 ~~in press~~). Sex differences in Learning Mathematics: A longitudinal Study ~~Using~~ with Item and Error Analyses. J. Educational Psychology, 79, 372-383.

R. E. Mayer (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. Instructional Science, 10, 135-175.

McCarthy, K.A. (1976). Sex Bias in Tests of Mathematical Aptitude. Dissertation submitted to the City University of New York. (University Microfilms, 76-11, 629).

Milton, G.A. (1958). Five studies of the relation between sex role identification and achievement in problem-solving. Technical Report No. 3, Department of Industrial Engineering and Department of Psychology, Yale University, New Haven, CT.

National Assessment of Educational Progress (NAEP). 1983. The Third National Mathematics Assessment: Results, trends, and issues (13-MA-01). Denver, CO: Education Commission of the States.

Paige, J.M. & Simon, H.A. (1966). Cognitive processes in solving algebra word problems. In B. Kleinmutitz (Ed.) Problem solving: Research, method, and theory (pp. 51-119). New York: John Wiley.

Polson, P.G. & Jeffries, R. (1985). Instruction in general problem-solving skills: An analysis of four approaches. In J. Segal, S. Chipman and R. Glaser (Eds.). Thinking and Learning Skills: Volume 1, (p. 417-458).

Rich, B. (1973). Modern Elementary Algebra: Schaum's Outline Series. New York: McGraw-Hill.

Riley, M., Greeno, J., & Heller, J. (1983). Development of Children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.). The Development of Mathematical Thinking. New York: Academic Press.

Segal, J.W., Chipman, S.F. & Glaser, R., (1985). Thinking and Learning Skills. Volume 1: Relating Instruction to Research. Hillsdale, NJ: Erlbaum.

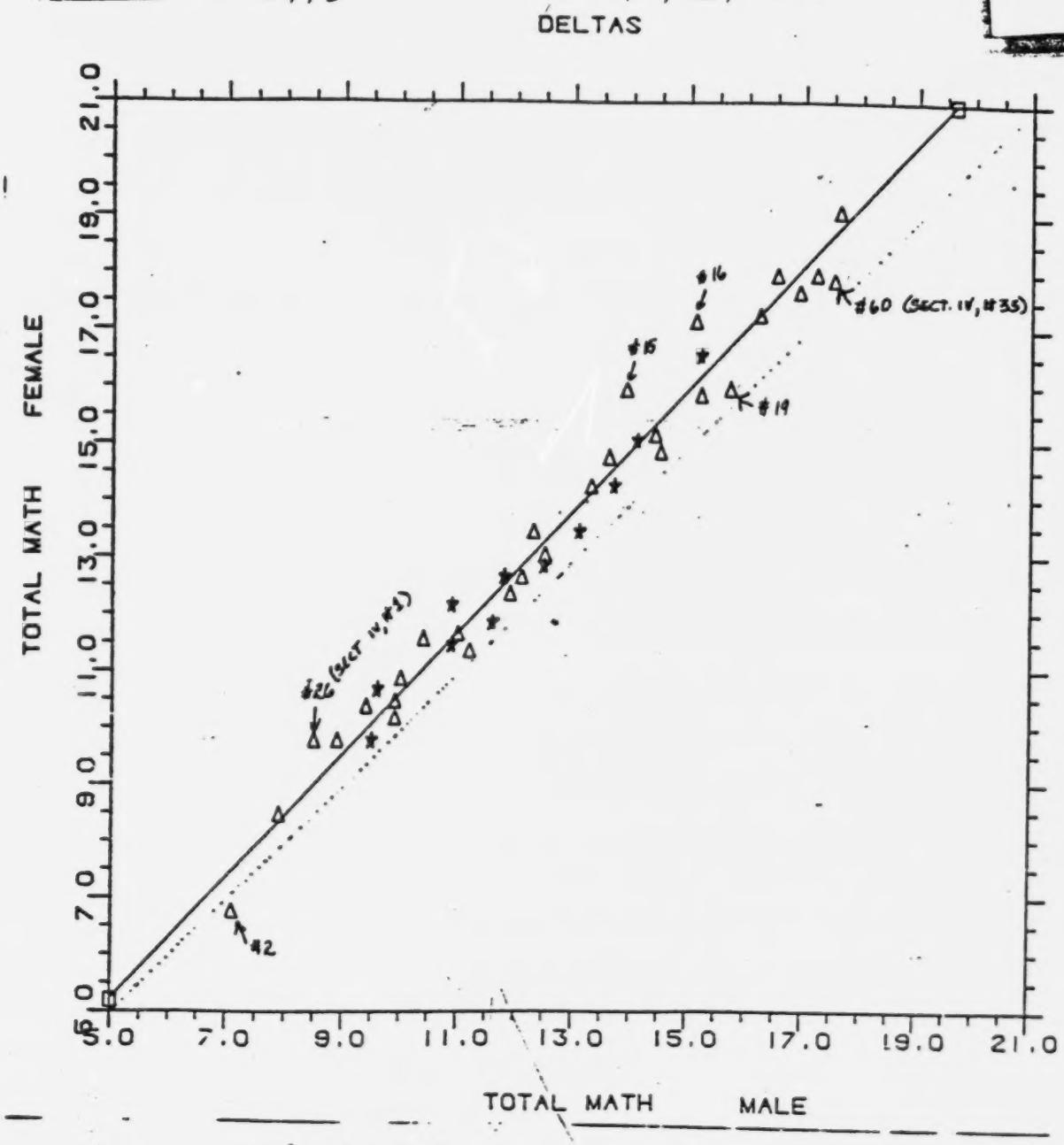
Silver, E.A. (1979). Student perceptions of relatedness among mathematical verbal problems. Journal for Research in Mathematics Education, 10, 195-210.

Strassberg-Rosenberg, B. & Donlon, T.P. (1975). Content influences on sex differences in performance on aptitude tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education. (ED. 110 493; TM 004 766.

VanDijk, T.A., & Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

Webb, N.L. (1979). Processes, conceptual knowledge, and mathematical problem-solving ability. Journal for Research in Mathematics Education, 10, 83-93.

Figure 2
from Strassenberg-Rosenberg & Donlon (1975) Content Influences on Sex Differences in Performance on Aptitude Tests
ED 110 493 TM 004 766



(see definition of Δ on Figure 2)

Figure 2. Delta Plot of SAT Total Math Section (n = 60 items) for Male and Female Sampl

r = .9868

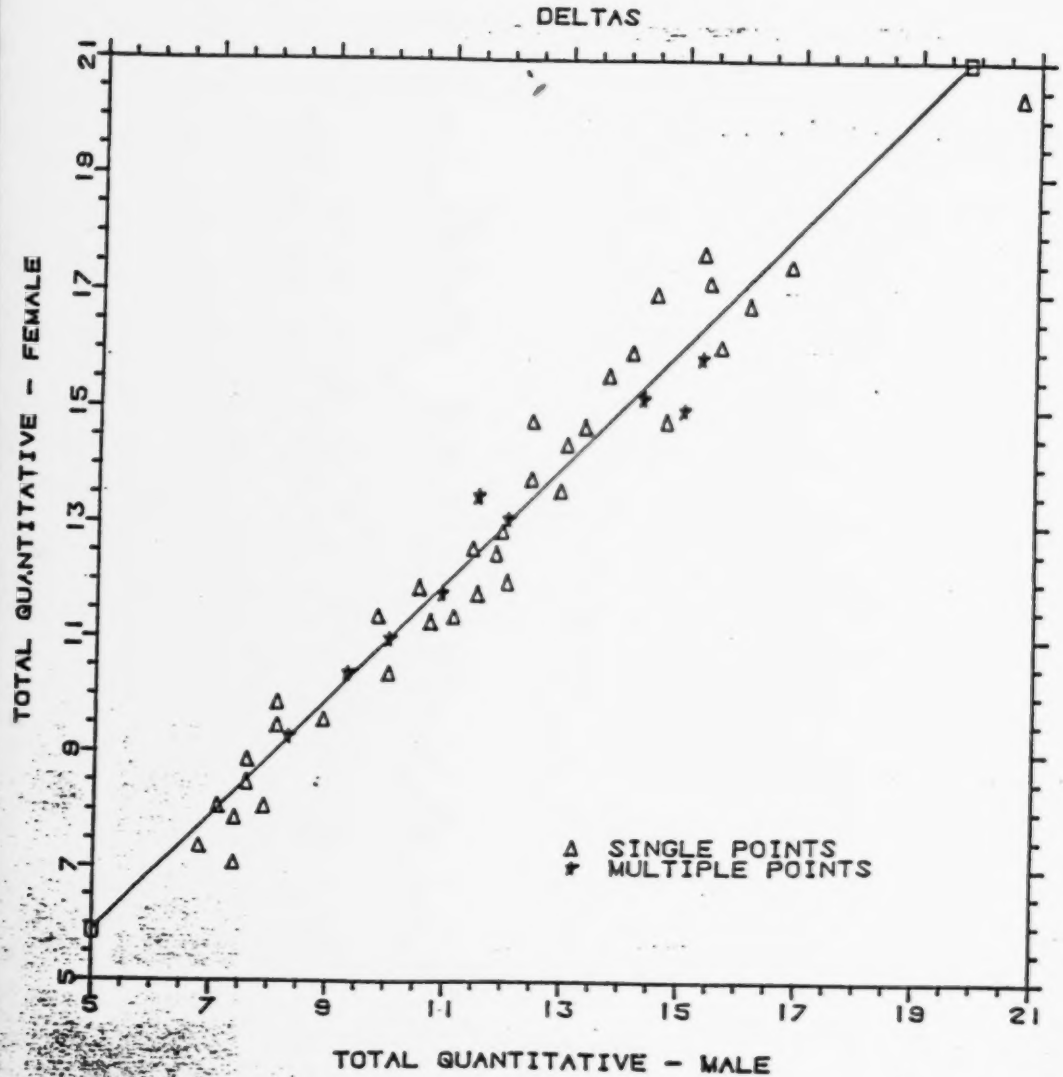
43

Δ Δ Δ SINGLE POINTS
★ ★ ★ MULTIPLE POINTS

Figure 2

from Donlon, Hicks, Wallmark, Sex Differences in Item Responses on the Graduate Record Examination (1980) Applied Psychological Measurement, 4, 9-20.

Figure 3
Δ Plot for Mathematics Section



Δ is defined as $\Delta = 4z + 13$ where z is a normal deviate above which lies a proportion of the area under the curve equal to the proportion of examinees answering the item correctly. The proportion p is based on

Table 1a

All effects are logits ± s.e.

exp.summ Fri Aug 7 09:02:49 1987

PROBLEM 08

year:	1986		1985	
topic:	lemonade	motoroil	lemonade	motoroil
resp:				
error	12	37	29	37
correct	255	193	209	151

For 1986, compared with 1985, large improvement in lemonade vers. smaller one in motoroil vers.

topic effect: 1.41 in 1986 (±0.35)
0.57 in 1985 (±0.27)
0.84 interaction
± 0.43 i.e., marginally signif.

PROBLEM 12 (5 Life Policies for every 4 Accident)

year:	1986		1985	
type:	numer	algeb	numer	algeb
resp:				
error	56	119	44	125
correct	218	104	158	99

type effect: 1.50 ± 0.14 (numeric > algebraic)
no effect of year (.09 ± .14)
no interaction (.02 ± .29)

PROBLEM 17

year:	1986				1985			
topic:	air		dairy		air		dairy	
type:	numer	algeb	numer	algeb	numer	algeb	numer	algeb
resp:								
error	33	73	8	22	57	84	11	25
correct	118	43	115	85	56	41	78	74

topic effect: 1.80 ± 0.16 assuming no interaction with year or type, which is tenable.

type effect: 1.69, 1986
0.74, 1985

type x year interaction: 0.95 ± 0.32

Interaction is due to large decrease in numerical errors on the air version, 32% in 1986, 50% in 1985.

PROBLEM 18

year:	1986				1985			
topic:	chair		nasa		chair		nasa	
type:	algeb	numer	algeb	numer	algeb	numer	algeb	numer
resp:								
error	108	90	105	85	94	92	64	71
correct	43	26	18	22	19	33	25	28
% correct	28.5%	22.4%	14.6%	20.6%	16.8%	26.4%	28.1%	28.3%

essentially no effects -- 23% correct for all cond.tions fits the data well

PROBLEM 19

year:	1986				1985			
topic:	pest		hem		pest		hem	
type:	algeb	numer	algeb	numer	algeb	numer	algeb	numer
resp:								
error	63	22	55	31	56	39	44	26
correct	88	94	68	76	57	86	45	73

type effect: 0.88 ± 0.14 numeric > algebraic
negligible topic effects
small year effect (0.27 ± 0.14)
and no significant interactions

— Large topic effects in 08 and 17, in the former case larger in 1986 than 1985
Little or no topic effect in 18 and 19

— Large type effects in 12, 17, 19; for 17, larger in 1986 than 1985
Little or no type effect in 18

— Not easy to interpret this

Chipman

Table 1b

8. James' outboard motor requires 1 1/2 quarts of oil per gallon of gasoline. How many gallons of gasoline will he have used when he has used twelve quarts of oil?

8. Mary's lemonade recipe calls for one and one/half lemons per pint. After making how many pints will she use up twelve lemons?

17. Kay goes to a dairy store and finds that Swiss cheese costs c cents per ounce while eggs cost g cents per dozen. She has enough money to buy d dozen eggs. If she decides to buy Swiss cheese instead of eggs, how many ounces can she buy?

17. A certain type of airplane can cover c kilometers per kilogram of fuel at low altitude but g kilometers per kilogram flying at high altitude. It carries enough fuel to travel a total distance of d kilometers at the low altitude. How many kilometers can it travel at the high altitude?

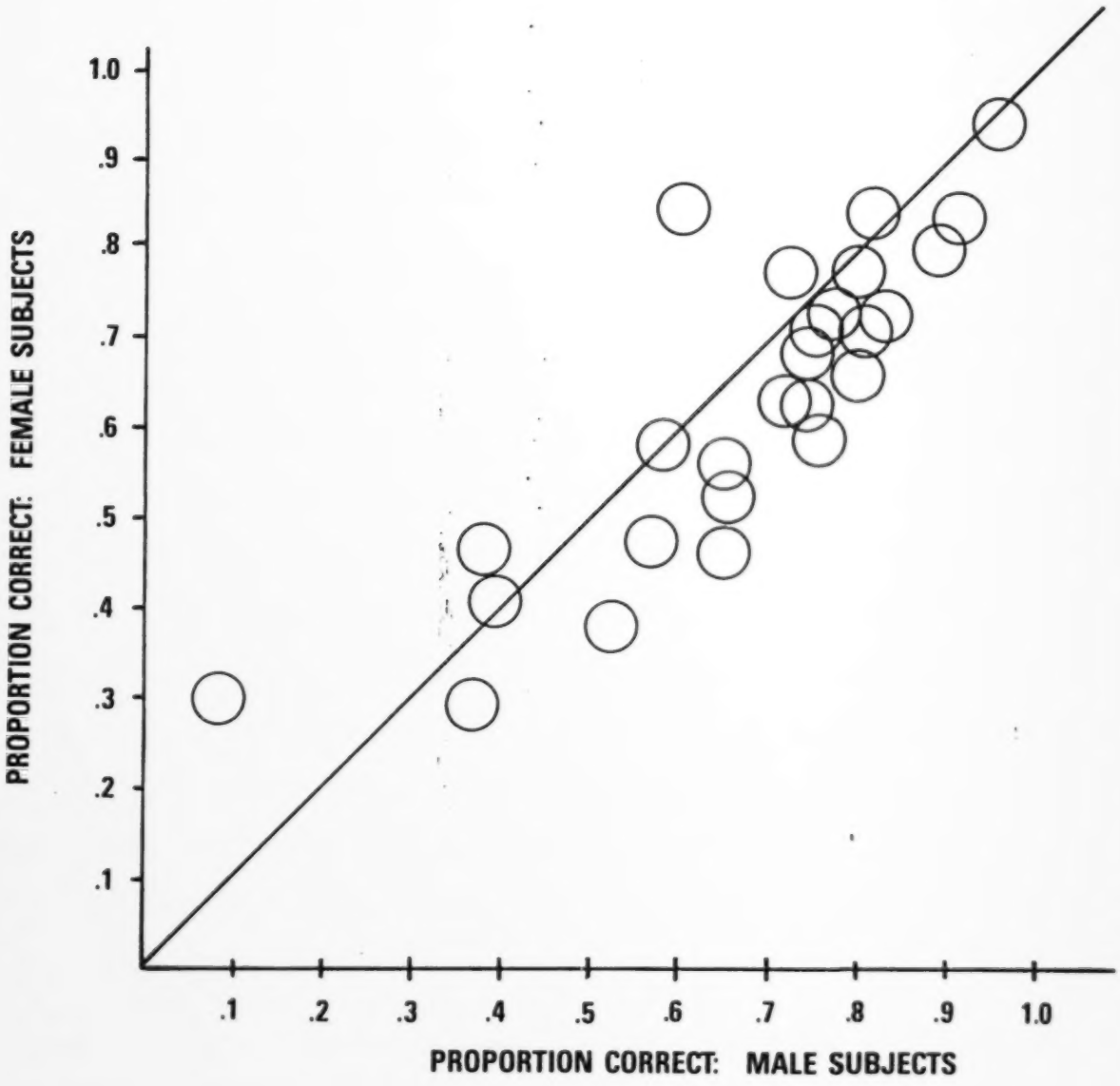
18. In order to cut costs for a mission, NASA has decided to shorten it from 27 days to 8 days. One result of this is that proportionately less oxygen will be needed on the flight. To reduce the size of the oxygen tank appropriately, by what factor should the designers multiply the lengths of the edges of the flat pieces of stainless steel that are put together to manufacture the tank?

18. Susan has a sewing pattern for a large bean-bag chair. The instructions call for 27 pounds of styrofoam beads. She learns that the beads are expensive and are sold only in cartons containing 8 pounds. She decides to make a smaller chair using only 8 pounds (one carton) of beads. By what factor should she multiply the lengths of the cuts for the pattern pieces in order to make a chair that requires only 8 pounds of beads?

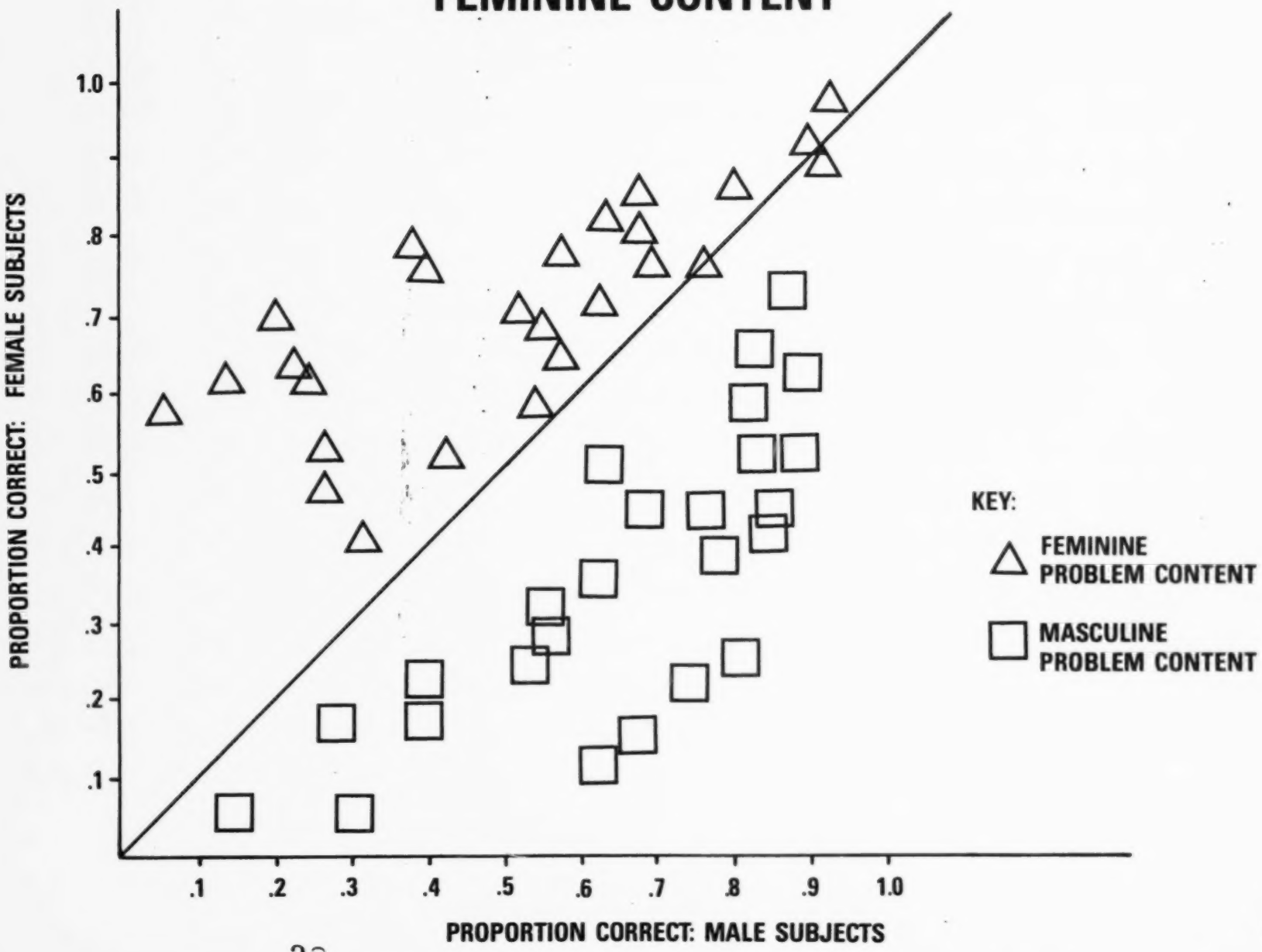
19. Molly and Tricia are responsible for hemming a set of drapes. Working alone, Molly could do it in 3 hours and Tricia could do it in 5 hours. Suppose they tackle the job together, each working at her usual rate on separate parts of the drapes. Which of the following expressions is equal to the proportion of the total job they can complete in one hour?

19. Two companies, No Pest Inc. and The Bugless Corp., each specialize in aerial spraying of insecticide on orchards. There is a certain large orchard that NoPest, with its turbo-charged planes and heavy-duty delivery system, can spray in 3 hours and Bugless (which has older equipment and charges less per hour) can spray in 5 hours. Suppose they tackle the job together, each working at its usual rate on separate parts of the orchard. Which of the following expressions is equal to the proportion of the orchard that they can spray in one hour?

EFFECT OF PROBLEM CONTENT: NEUTRAL CONTENT



EFFECTS OF PROBLEM CONTENT: MASCULINE AND FEMININE CONTENT



Chipman

Table 2

23. The batter which is 7 inches high in a loaf pan 1-1/4 feet by 8 inches, is poured into a pan 13 inches by 20 inches. What height will it reach in the larger pan?

32. A secretary types business letters at a rate of d words per minute, and statistical tables at a rate of a words per minute. What is her average rate of typing if she types for 2 hours?

Female outliers

41. Mrs. Crocker can do the dinner dishes in a quarter of an hour, while her daughter takes half an hour. What part of an hour would it take to do the dishes if they work together?

74. If two cups of melted chocolate weigh 16 ounces, how many cups of melted chocolate can be obtained from a package weighing 1 pound 8 ounces?

113. Barbara spends $11.22 purchasing and mailing party invitations. If the invitations are purchased for 3 for 10¢, and each requires an 8¢ stamp, how many invitations were purchased?

25. Jane cut a length of cloth 13 yards, 5 feet, and 1 inch into three equal parts. The length of each piece is?

8. In making drapes, Joan needs to cut 8 foot panels. How many panels may she cut from fabric that is y yards long?

Extreme male outliers

15. A mile runner is clocked at 58 seconds after the first quarter mile (440 yards), and 1 minute 56 seconds after he has run a half mile. What rate (in yards per second) must he maintain for the final two quarters if he is to run a 4 minute mile?

44. Ten merchants agreed to purchase uniforms for a local baseball team at a total cost of m dollars. After a disagreement, 2 merchants dropped out of the project. By how many dollars was the cost to each remaining merchant increased?

116. A trackman runs 100 yards in 9.8 seconds and then runs 440 yards in 49 seconds. What is the ratio of his average speed in the 100 yard run?

66. The indicator of an oil tank shows 1/5 full. After the truck delivers 165 gallons of oil the indicator shows 4/5 full. What is the capacity of the tank (in gallons)?

---

85. Because of a decrease in available man power, a tool factory reduced its monthly output by 20%. What was the necessary percent increase of man power needed to bring the output up to normal production?

Other Male Outliers

16. A carpenter is supposed to build a minimum of t tables. In d days he builds m more tables than the minimum. What is the average number of tables he produced each day?

6. One gear rotates once every 7 minutes, and another rotates once every 5 minutes. Both gears begin to rotate at the same time every (?) minutes.

54. Seven type p nails are worth 5 type q nails. Three type p nails are worth x type q nails. What is the numerical value of x?

40. A motorist leaves at 9:00 AM and stops at 9:20 AM. If he covered 18 miles, what was his average velocity for this part of the trip?